Performance of Molecular Inversion Probes (MIP) in Allele Copy Number Determination

Yuker Wang*[1], Martin Moorhead*[1], George Karlin-Neumann[1], Nicholas J. Wang[2], James Ireland[1], Steven Lin[1], Chunnuan Chen[1], Laura M Heiser[2], Koei Chin[3], Laura Esserman[3], Joe W. Gray[2], Paul T. Spellman[2$], and Malek Faham[1$]

1 Affymetrix Inc. 7300 Shoreline Blvd South San Francisco CA 94080

2 LBL 1 Cyclotron Rd, MS977R225A, Berkeley CA 94720

3 Comprehensive Cancer Center, 2340 Sutter Street, University of California San Francisco, San Francisco, CA 94143

* these authors contributed equally

$ Correspondence should be addressed to these authors

Paul Spellman
LBL 1 Cyclotron Rd, MS977R225A, Berkeley CA 94720
PTSpellman@lbl.gov

Malek Faham
Affymetrix Inc. 7300 Shoreline Blvd South San Francisco CA 94080
Malek_faham@affymetrix.com

Abstract

We have developed a new protocol for using Molecular Inversion Probes (MIP) to accurately and specifically measure allele copy number (ACN). The new protocol provides for significant improvements including the reduction of input DNA (from 2μg) by more than 25 fold (to 75ng total genomic DNA), higher overall precision resulting in one order of magnitude lower false positive rate, and greater dynamic range with accurate absolute copy number up to 60 copies.

Background

Chromosomal copy number analysis has been important in the study of tumor samples for decades. Changes in copy number have already been demonstrated to predict patients' response and/or prognosis, [1] which gives hope that this can be applied in large scale to significantly affect clinical care in the future. In order to fulfill this promise, technologies that are able to assess copy number on the whole genome scale in a large number of samples are required. Since the development of Comparative Genomic Hybridization (CGH), [2] many technologies have been developed to address this need. These include Bacterial Artificial Chromosome (BAC) CGH and more recently CGH employing several types of oligonucleotides arrays [3-7]. Some of the newer CGH methodologies allow for allelic information to be obtained [4, 5, 7, 8]. The utility of measurement of allele copy number include the identification of loss of heterozygosity (LOH) events [4] and the allelic composition at amplified loci [9].

One of the techniques that have previously been described for the measurement of allele copy number is molecular inversion probes (MIP) [10, 11]. Briefly MIP probes are circularizable oligonucleotides, where the two ends carry two sequences that are complementary to two sequences on the genome separated by one nucleotide (exactly where the variant to be genotyped is). After hybridization to the genomic DNA, the reaction is split into 4 tubes where a single nucleotide is added to each tube. Upon the addition of the nucleotide (only in the tube with the nucleotide that is complementary to the allele on the genome) the MIP probe is then ligated turning the probe into a circle. This structure can be selected for by the use of exonucleases allowing for minimal "cross talk" between probes and the ability to obtain high quality data from highly multiplexed assays (>50,000 plex). Ultimately these products are amplified and hybridized onto an Affymetrix microarray to identify the present products.

The MIP assay differs from other highly multiplexed (10,000s-100,000s) techniques in that it utilizes enzymatic steps in solution to capture specific loci, which is then followed by an amplification step. Such combination of enzymatic steps confers a high degree of specificity on the MIP assay.High specificity and minimum "cross talk" between loci or alleles results in precise measurements as well as large assay dynamic range. In addition the amplification of the loci of interest only simplifies the task of detection and provides the ability to use lower amounts of input genomic DNA. The high precision, large dynamic range, and low DNA usage are demonstrated in this study. Finally, because MIP requires only 40 bp of intact genomic DNA, its use in degraded samples, like Formaldehyde Fixed Paraffin Embedded (FFPE) may offer distinct advantages.

We have made significant advancements in this technology. As a result, the false positive rate decreased by an order of magnitude and the dynamic range extended to achieve accurate absolute copy number measurements up to 60 copies, while reducing the input genomic DNA requirement by more than 25 folds.

We describe the performance of the MIP assay using several types of metrics that are broadly useful to all copy number assays: (1) the ability to discriminate a copy number aberration from normal at a total as well as allelic copy number level, and (2) the ability to accurately quantitate the level of copy number aberration at both total and allelic copy number levels.

Results

MIP copy number assay modification

We have previously described the use of MIP for copy number analysis [10]. We have now improved the performance of the technology through modifications of the MIP copy number protocol and through improved data analysis. The improved performance allows allele copy number data to be obtained using 75ng of human genomic DNA.

The first implementation of the MIP ACN assay required 2μg of genomic DNA. We discovered that only a fraction of the genomic templates hybridized to MIP probes that are then circularized and amplified. We hypothesized that increasing the number of MIP molecules and decreasing the hybridization volume should increase the number of MIP molecules bound to their genomic targets. We tested this hypothesis and verified that increasing the number of MIP molecules by a factor of four and decreasing the hybridization volume (from 27 μl to 6.7 μl) allowed us to substantially decrease genomic DNA input. After the hybridization, buffer is added to increase the volume to 27 μl, and the rest of the protocol is unmodified.

In the standard genotyping protocol genomic target is split into 4 reactions, where one of each of the 4 nucleotides is added. We recognized that we could decrease DNA input requirements by performing a smaller number of these reactions We reasoned that if we were to only use one set of SNPs (for example only the most common C/T SNPs), we would decrease the DNA requirement by 50%. Similarly, adding two nucleotides into each of two reactions leads to the same result. We have implemented this variant protocol by adding G and C nucleotides into one tube, and adding A and T into another. In this scenario, about 85% of SNPs in the human genome (all but G/C and A/T SNPs) can be assessed. An advantage of decreasing the number of reactions is that it requires only two independent readouts rather than four (i.e., 4 colors on 1 array or 1 color on 4 arrays). In the optimized procedure, 75 ng genomic DNA is mixed with more than 50,000 probes in a small volume (6.7 μl). The hybridized probe:target genomic DNA are split into two reactions, where 2 nucleotides are added to each of the two tubes. The two

reactions are processed separately and read on two independent arrays, which was found to yield better data than two colors on one array (data not shown).

One effect that requires correction in quantitative assays on arrays is the phenomenon of saturation. This is especially important for correct estimations of amplifications. We have implemented a Langmuir correction for the non-linear relationship between signal and copy number [15]. Our algorithm was developed on a separate data set, and the data shown here is an independent set. Using this algorithm we have been able to measure copy number in a linear fashion at levels over 60 copies (see below).

Detection of aberrations:

An important aspect of the copy number performance is the detection of aberrations where the copy number is distinct from 2. The degree of discrimination between copy number 2 and the aberrant copy can be understood through ROC curves showing the trade off between false positive and sensitivity (1-false negative rate) given data on regions with known copy number. The presence of cell lines carrying 1,3,4, or 5 X chromosomes provides a good resource for the study of the performance of the technology in this copy number range [2]. For example in the assessment of the cell lines with one X chromosome (males) one can make a threshold at copy number 1.5 and any marker on the X chromosome with a copy number below 1.5 would be considered a true positive, and any autosomal marker with a copy number below 1.5 is considered a false positive. By plotting this trade off between true and false positives at many thresholds between copy numbers of 0 to 3 the full ROC curve is generated.

To assess the ability of MIP to detect copy number aberrations we used a probe panel containing ~53,000 single nucleotide polymorphisms (SNP). We utilized this pool to assay 63 samples (45 unique, 9 duplicate) from the 3 major populations used in the HapMap project. Out of the 53,341 SNPs, 50,806 had genotyping call rates of greater than 90%. We then sorted the remaining SNPs based on the standard deviation of their predicted copy number. We selected the most robust markers for detailed study of copy number performance by selecting those with a standard deviation of less than 12%. This yielded a population of 39,785 markers. Figure 1 shows the copy number estimates across the genome for the different samples carrying 1-5 copies of the X chromosome. By assuming that males have only one copy of the X chromosome markers and two copies of autosomal markers, we generated ROC curves to describe the trade off between false positive and sensitivity for distinguishing one copy from two copies (Figure 2, red line). Similar ROC curves can be generated for the discrimination between 2 and 3, 4, or 5 copies (Figure 2). By comparing the generated ROC curves with our published data for previous MIP protocol we find a dramatic improvement. For example, at the same 50% sensitivity level, we found a reduction of the false positive rate by an order of magnitude.

The ROC curve above describes the average performance of a set of samples. We also wished to understand the performance of individual samples. As can be seen in Figure 3, individual samples have different false positive rates given the same sensitivity level.

Similarly ROC curves can be generated to assess the ability to study allele copy number. For example Figure 4 depicts the ROC curve to assess the ability to discriminate the usual 1:1 ratio in heterozygotes from 2:1 ratio on the X chromosome in a cell line carrying 3X chromosomes. The ROC curve for allele ratio is not as good: at a sensitivity

level of 50%, the copy number false positive rate is ~ $1x10^{-3}$, and the allele ratio false positive rate is ~$8x10^{-3}$. One reason for this discrepancy is that we are using the best markers as defined by copy number root square deviation (RSD). The use of the best markers as defined by an allele ratio criteria (allele ratio RSD) significantly improves the performance (sensitivity of 50% and false positive rate of ~$3x10^{-3}$

Systematic false positives

The above analysis assumes that all the autosomal markers are present at two copies per cell. There has been a wealth of evidence demonstrating copy number polymorphisms (CNP) in the general population [12, 13]. Therefore a fraction of what we considered as false positive may in fact be true positives. In addition, the presence of a secondary SNP (distinct from the one being interrogated) within the probe may emulate the presence of a deletion. Data generated on two CEPH pedigree populations, Yoruban and Utah, is informative in this regard because the polymorphism on which the MIP panel are based are from European (equivalent to Utah) rather than African populations. The contribution of genetic variants (CNP or SNP) to the apparent false positive rate is suggested by our detection of ~ 3 fold more apparent autosomal deletions in the Yoruban population compared to the Utah population (average number of markers per sample with measured copy number below 1.3 is 126 markers for Utah population and 319 for Yoruban population). We hypothesized that this imbalance between the number of apparent deletions in the two populations was likely due to secondary polymorphisms close to the SNP being assayed which prevent proper binding of the MIP to its target. Further evidence to support this hypothesis was noted when we observed that the majority of these apparent deletions were reproducible when a sample is re-assayed.

To understand the nature of these apparent deletions we randomly picked ten SNPs, which showed copy number measurements below 1.3 in replicate measurements from the Yoruba sample (sample NA18515). We PCR amplified ~400 basepair fragments that included the SNP assayed by MIP and used dideoxy sequencing to show that eight of these nine loci that were successfully sequenced had a secondary SNP within the MIP probe homology sequence. The ninth SNP that showed copy number 1 was assayed by qPCR to measure copy number but was found to show a normal copy number of two (Supplementary Table 1).

Trade off between resolution and performance

Copy number changes are expected to occur in discrete segments allowing neighboring markers to be averaged together. This leads to enhanced performance as measured by trade off between false positive and sensitivity (i.e. ROC curve moving to the upper left) at the expense of lower resolution.

As discussed above, one shortcoming of the ROC analysis is the presence of CNPs in the autosomes. Averaging two adjacent markers that lie within a CNP will erroneously consider these markers as false. Therefore for the purpose of describing the performance of the technology we averaged markers that are not adjacent for each other. This method would ameliorate the effect of miscalling two adjacent markers in a CNP as a false positive. This analysis is appropriate as long as there is a lack of correlation between marker performance and the position on the chromosome. If this assumption is true then the operation reflects the performance of averaging two adjacent markers since the

adjacent and the random markers are obtained from the same distribution. Clearly averaging data from non-adjacent markers is only valid for the assessment of the technology performance and cannot generate any meaningful biological findings.

Averaging over two markers improves the performance of the MIP data significantly (Figure 5). Clearly when one is trying to obtain biological information smoothing non-adjacent markers is totally erroneous. In this case we were interested in the exact opposite: erasing any real biological information (copy number polymorphisms) and hence we smoothed across non-adjacent markers. For the discrimination between 1 and 2 copies, a sensitivity level of 80% and a false positive rate of $5x10^{-5}$ can be achieved.

The ROC curves shown in the above figures describe the performance of the top ~75% of the markers in the panel we constructed. It is expected that as more of the lower quality markers are considered, the ROC performance will decrease. We included ~48,000 markers (~90% of the total) in the analysis. Figure 5 shows the ROC curve to discriminate one from two copies using one marker or two markers using 75% (40K) or 90% (48K) of the data. As can be seen in Table 1 the average performance with 90% of the markers is somewhat worse than that seen with 75% of the markers when judging the specificity at 50% sensitivity.

Accuracy of copy number estimation

The ROC curves describe the discrimination between two copies and a specific aberration. However it does not define the accuracy of the copy number estimation. The accuracy of the copy number determination can be estimated by the deviation from the true copy number. This can be readily measured for 1-5 copies using the X chromosome series. As can be seen in Table 2, the copy number estimation in the MIP data is very close to the true value. The precision, as defined by the relative standard deviation, over the 1-5 copy number range is 0.1-0.14.

Accuracy of copy number estimation at high copy number amplification can be assessed by comparing the MIP estimation with real time PCR measurement. We have done such a calibration for a selected amplification in cell line MCF7 (Figure 6). The average copy number estimate among 30 MIP markers within the amplification is 43, which is close to the 33 copies measured by real time PCR. Copy number estimation is computed relative to a "control" region in the genome. In cancer cell lines, the "control" region used in real time PCR may not have the average ploidy of the cell and therefore may bias the estimation of the amplified region. In fact in this example the control region was from chromosome 2, which is estimated to be present at slightly elevated copy numbers compared to the average of the genome based on the MIP data. Correcting for this bias would make the MIP and real time PCR copy number estimation of the amplification even closer.

To carefully assess the accuracy of the measurement at high copy number values, we added a known quantity of a set of PCR amplicons to a normal sample before the MIP reaction was performed. The DNA fragments that were spiked in were added at different copy number levels ranging from no extra copies to several hundred additional copies. Supplemental Table 2 shows the PCR amplicons, the MIP probes they correspond to, and the spike in levels. We show the relationship between the expected and the measured copy number of either the individual spikes (Figure 7).

Accuracy of measurement of allele copy number in amplification sites for many methods is limited by allele cross talk. Allele cross talk is the proportion of signal measured for one allele in the presence of a second allele. To assess this phenomenon using MIPs we studied the spike in data. The spiked in PCR amplicons were purposely generated from an individual that is homozygous and added into DNA from a heterozygous individual making the copy number for one allele 1 and the other ranges from 1 to 1,000. The allele cross talk in the MIP assay is very low, as the presence of 100 copies or more of one allele does not change the copy number of the other allele significantly (Table 3).

Identification of LOH without Matched Normal Tissue

A major challenge in the study of ACN is the absence of matched normal tissue for many valuable clinical samples. In tumors that have lost one allele, it is not easy to discriminate LOH for individual alleles that are homozygous in the entire individual. We recognized that the high sensitivity and accuracy of the MIP ACN assay, coupled with the high likelihood of normal tumor contamination, could allow us to distinguish LOH from alleles that are homozygous. In theory, this should be best accomplished with tumor showing substantial (approaching 50%) normal contamination.

To test this theory, we analyzed ACN from 5 breast tumors using the 60K MIP panel. Visual examination of the data clearly show a typical plot of estimated copy number for allele A vs. allele B, compared to a tumor with relatively normal genome structure (Figure 8A). Three clusters are expected in such a plot, one at ~2,0 (homozygous A), one at 0, ~2 (homozygous B), and one at ~1, ~1 (heterozygous). In the aberrant tumor samples (Figure 8B, C), three distinct clusters can be observed in the heterozygous cluster. The central cluster represents the "true" heterozygous copy number measurements. The flanking clusters represent LOH of either the A or B allele. These sub-clusters of the heterozygous cluster clearly resolve into discrete copy number segments along the chromosome as can be seen in Figure 9. We are also able to observe that deletions are observed not as zero copies for each allele, but as ~0.5 copies of each allele (Figure 9D). To assess reproducibility, we analyzed all samples in duplicate and calculated concordance estimates for the various genotypes (Table 4).

Discussion

We describe in this manuscript significant improvements we have made to the MIP-based measurements of ACN. By increasing the proportion of genomic targets that are hybridized to the MIP probes, we have improved the performance while requiring a smaller amount of DNA. Additionally, for copy number measurements there are substantial advantages in uniformity and robustness when utilizing one-color readouts, especially at high levels of multiplexing. The use of a control sample that is co-hybridized with the test sample in an analogous fashion as used by BAC arrays leads to inferior results compared with the one color readout (data not shown). Presumably this is because the different dyes have different characteristics of brightness and saturation. We conclude that the effect of the lack of uniformity among the dyes is probably larger in our system than chip-to-chip variation that the control sample co-hybridization is supposed to ameliorate. The improvements achieved from the new protocol as evaluated by ROC

curve analysis resulted in a decline in the false positive rate by an order of magnitude, while reducing the input genomic DNA by more than 25 fold. In addition the dynamic range has been extended with accurate estimation achieved for up to 60 copies.

We evaluated the performance of MIP for ACN measurements using a set of metrics that are broadly useful for all copy number assays. We demonstrate the ability of MIP to detect a single copy deletion or duplication at allele and total copy number using ROC curve analysis. We believe ROC curve analysis provides a rigorous statistical framework for comparing different technologies or different protocols/algorithms of the same fundamental technology. In addition to genuinely improving the technology performance in the ROC curves by the use of better protocol and algorithms, one may apparently improve them by smoothing (figure 5), filtering the worst markers (figure 5), or the worst samples (figure 3).

We have shown in the single MIP marker analysis that many of the apparent false positives in the discrimination between 1 and 2 copies are due to the presence of SNPs in the genomic sequence that are complementary to the MIP probes. This effect will be strongest in the populations that are most diverse. It should be possible to ameliorate this effect by using matched normal and tumor pairs. The presence of SNPs may explain why the discrimination between 1 and 2 is not better than that between 2 and 3, as secondary SNPs that interfere with MIP binding emulate a copy number deletion.

We also show the MIP assay precision of measurements of copy number at allele and total copy number. Precision at the total copy number requires low background of the assay and lack of saturation. In addition the allele level precision requires low level of allele cross talk even when one allele is present in huge excess relative to the other.

These observations led us to suspect that it should be possible to genotype mixed DNA populations, such as occurs in tumor samples contaminated with normal tissue. As normal contamination increases some estimate of the amount of normal contamination is valuable, which we believe can be quite accurately estimated using the calculated copy numbers for regions of LOH and deletion.

One promise of ACN data over the traditional total copy number data is the potential that it may facilitate the identification of the critical genes in regions of aberrations. Even though large aberrations can be readily identified by total copy number CGH, the identification of the critical gene(s) in these aberrations is often not straightforward. This is in contrast to sequencing data where identification of mutations has been quite laborious, but once achieved the critical gene is usually easily identified. Identification of an allele that is preferentially deleted or amplified in a set of samples implicates the specific allele (or one in linkage disequilibrium with it) as critical in the pathogenesis of the aberrations. The one example in the literature may be supplemented by assessment of allele copy number in a large number of SNPs in different sets of cancer samples.

Material and Methods

Samples and MIP assay:

The normal samples as well as the samples carrying 3 (NA04626), 4 (NA01416), and 5(NA06061) copies of the X chromosome were obtained from Coriell Cell Repository (Camden, NJ). The normal HapMap samples that were used were also obtained from Coriell Cell Repository. The samples that were used were: NA19240, NA19239, NA06991, NA06985, NA19238, NA19222, NA19202, NA19201, NA19200, NA19132, NA19131, NA18956, NA18951, NA18949, NA18947, NA18945, NA18912, NA18854, NA19130, NA19128, NA19127, NA19099, NA19094, NA18991, NA18987, NA18981, NA18605, NA18603, NA18582, NA18573, NA18558, NA18550, NA18547, NA18542, NA18537, NA18515, NA18508, NA12892, NA12813, NA12717, NA12156, NA12155, NA12004, NA11881, NA11840, NA11832, NA11830, NA10831, NA07345, NA07056, NA07029, NA07019, NA07000, and NA06993. The MCF7 cell line was obtained from the American Tissue Cell Culture (ATCC).

The MIP assay was performed as described previously, but with important modifications [14]. Specifically, the current protocol is a modification of the Targeted Genotyping protocol commercialized by Affymetrix (Additional information about MIP technology can be found at the Affymetrix website, http://www.affymetrix.com/products/application/targeted_genotyping.affx). Briefly, test DNA samples were diluted to 16ng /μl. All DNA quantitation is done using PicoGreen dsDNA Assay Kit (Molecular Probes / Invitrogen, P7589). 96 or 384 well plates were used whenever possible to reduce variation. For day1 overnight annealing, 4.7 ul of DNA samples (75ng total), 0.75ul of Buffer A, 1.1 ul of the 53K probe pool (200 amol/ul/probe) and 0.045 ul of Enzyme A were mixed well in a 384-well plate on ice. The reaction was incubated at 20°C for 4 min, 95°C 5 min, then 58°C overnight. On day2, 13 μl of Buffer A was added to each well with 1.25 μl of Gapfill Enzyme mix. 9 μl is taken to each of two wells in a 96-well plate. MIP probes were circularized by 4ul of di-nucleotide (dATP with dTTP, dCTP with dGTP) mix at 58°C for 10 min. The uncircularized probes and genomic DNA were eliminated by addition of 4 ul of Exonuclease Mix and incubation at 37°C for 15 min, followed by heat-killing of enzymes. The circularized probes were linearized by the addition of Cleavage Enzyme Mix at 37°C for 15 min, then subjected to universal primer amplification for 18 cycles at 95°C 20 sec, 64°C 40 sec and 72°C 10 sec. For the labeling reaction, the product was further amplified with the label primers for 10 cycles, and then subjected to cleavage by Digest Enzyme Mix at 37°C for 2 hours. To hybridize, the cleaved MIP products were mixed with hybridization cocktail, denatured and hybridized to 70K Universal Taq arrays at 39°C for 16 h (two arrays per sample). The overnight hybridized arrays were washed on GeneChip® Fluidics Station FS450 and stained by SAPE at 5ng/ml (Invitrogen).

Copy number estimation was obtained from the hybridization signals as described previously, but with the following modifications [10]. Given that in this work no multi-color readout was present (but rather single color readout on two arrays), no spectral overlap was present, therefore the color-seperation step was omitted. In addition instead of the linear calibration of the allele signals, Langmuir correction was done [15].

Generation of Spike-In Samples

A panel of 80 PCR products representing genomic regions containing MIPs on chromosome 2 were PCR amplified from CEPH1341.14 (NA06985) using an ABI 9700 thermocycler (initial denaturation of 95°C for 5 minutes, 95°C for 30 s, 58°C for 30 s, 72°C for 60 s for 30 cycles; final extension at 72°C for 7 minutes). The products were purified using a MinElute 96 UF PCR Purification plate (Qiagen) and resuspended in TE. The purified products were quantitated on a fluorometer using the Quant-It™ dsDNA Assay Kit (Invitrogen). Purified PCR products were then pooled into 10 tubes, each containing 8 different products (supplementary Table 2). Each pooled tube of probes was then serially diluted 2-fold into a series of spike-in tubes containing 150 ng of genomic DNA from CEPH1341.02 (NA06991) (supplementary Table 2). The genomic DNA samples were chosen so that the spike-in PCR products from CEPH1341.14 represented a single allele, while the genomic DNA from CEPH 1341.02 was heterozygous, allowing for discrimination of allele specific amplification.

Sequence Analysis of Aberrant MIPs

PCR products were amplified using primers designed to span sequences containing MIPs that did not hybridize as expected (supplementary Table 3). Amplification was carried out in a 50 ul reaction (initial denaturation of 95°C for 5 minutes, 95°C for 30 s, 58°C for 30 s, 72°C for 60 s for 30 cycles; final extension at 72°C for 7 minutes) and products were purified using a MinElute 96 UF PCR Purification plate (Qiagen) and resuspended in TE. The purified products were sequenced using an Applied Biosystems 96 capillary 3730xl DNA Analyers and the forward and reverse primers used during amplification.

Identification of LOH without matched normal tissue

Genotyping metrics from the traditional MIP method were applied to each observation and estimated genotypes (AA, AB, or BB) were determined for each MIP in each of 5 replicated tumor samples. Data are provided as supplementary file 1. Regions of the genome that show clear evidence for decreases in copy number are easily observed with the decrease in copy number equivalent to 1.5 total copies, (1 copy of one allele and 0.5 copies of the other, or for homozygous alleles 1.5 total copies). No regions of the genome in any of the 5 samples analyzed appear to have ~1 copy of the higher allele and ~0 copies of the lower allele.

List of Abbreviations

ACN Allele Copy Number
MIP Molecular Inversion Probe
CGH Comparative Genome Hybridization
BAC Bacterial Artificial Chromosome
LOH Loss of Heterozygosity
SNP Single Nucleotide Polymorphism
CNP Copy Number Polymorphism
ROC Receiver Operator Characteristic

References:

1. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M *et al*: **Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2**. *N Engl J Med* 2001, **344**(11):783-792.
2. Kallioniemi OP, Kallioniemi A, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D: **Comparative genomic hybridization: a rapid new method for detecting and mapping DNA amplification in tumors**. *Semin Cancer Biol* 1993, **4**(1):41-46.
3. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J *et al*: **Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation**. *Genome Res* 2003, **13**(10):2291-2305.
4. Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR *et al*: **High-resolution analysis of DNA copy number using oligonucleotide microarrays**. *Genome Res* 2004, **14**(2):287-295.
5. Zhao X, Li C, Paez JG, Chin K, Janne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C *et al*: **An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays**. *Cancer Res* 2004, **64**(9):3060-3071.
6. Barrett MT, Scheffer A, Ben-Dor A, Sampas N, Lipson D, Kincaid R, Tsang P, Curry B, Baird K, Meltzer PS *et al*: **Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA**. *Proc Natl Acad Sci U S A* 2004, **101**(51):17765-17770.
7. Ishikawa S, Komura D, Tsuji S, Nishimura K, Yamamoto S, Panda B, Huang J, Fukayama M, Jones KW, Aburatani H: **Allelic dosage analysis with genotyping microarrays**. *Biochem Biophys Res Commun* 2005, **333**(4):1309-1314.
8. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J *et al*: **High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping**. *Genome Res* 2006, **16**(9):1136-1148.
9. Ewart-Toland A, Briassouli P, de Koning JP, Mao JH, Yuan J, Chan F, MacCarthy-Morrogh L, Ponder BA, Nagase H, Burn J *et al*: **Identification of Stk6/STK15 as a candidate low-penetrance tumor-susceptibility gene in mouse and human**. *Nat Genet* 2003, **34**(4):403-412.
10. Wang Y, Moorhead M, Karlin-Neumann G, Falkowski M, Chen C, Siddiqui F, Davis RW, Willis TW, Faham M: **Allele quantification using Molecular Inversion Probes (MIP)**. *Nucleic Acid Research* 2005, **28**:e183.
11. Ji H, Kumm J, Zhang M, Farnam K, Salari K, Faham M, Ford JM, Davis RW: **Molecular inversion probe analysis of gene copy alterations reveals distinct categories of colorectal carcinoma**. *Cancer Res* 2006, **66**(16):7910-7919.

12. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M *et al*: **Large-scale copy number polymorphism in the human genome**. *Science* 2004, **305**(5683):525-528.
13. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: **Detection of large-scale variation in the human genome**. *Nat Genet* 2004, **36**(9):949-951.
14. Hardenbol P, Yu F, Belmont J, Mackenzie J, Bruckner C, Brundage T, Boudreau A, Chow S, Eberle J, Erbilgin A *et al*: **Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay**. *Genome Res* 2005, **15**(2):269-275.
15. Burden CJ, Pittelkow YE, Wilson SR: **Statistical analysis of adsorption models for oligonucleotide microarrays**. *Stat Appl Genet Mol Biol* 2004, **3**:Article35.

Figure legends

Figure 1. Genomic view of samples with 1-5X chromosomes

The X axis shows the markers in a genomic order, with each chromosome uniquely colored. The Y chromosome depicts the measured copy number for each marker in linear scale. The X chromosome is the last shown chromosome in orange color. A) A male sample with 1 X chromosome. B) A female sample with 2X chromosomes. C) A cell line with 3 X chromosomes. D) A cell line with 4 X chromosomes. E) A cell line with 5 X chromosomes.

Figure 2. ROC analysis

The X axis is the rate of false positives (in $\log_{10}$), computed as the proportion of autosomal marker that have copy number below any given threshold (for the 1X calculation). The Y axis depicts sensitivity, defined as the proportion of X chromosome markers that have copy number values below the same threshold (for the 1X calculation). The curve is generated by calculating these values at many different thresholds. The curves from the 3X, 4X, and 5X cell lines were generated in an analogous fashion.

Figure 3. ROC analysis for individual samples

The X axis is generated in the same fashion as Figure 2 with the exception the curve for each sample is plotted separately. The average curve is the thick black line.

Figure 4. ROC analysis for allele ratio

The X axis is the rate of false positives (in $\log_{10}$), computed as the proportion of autosomal marker that have allele ratio above a threshold. The Y axis depicts sensitivity, defined as the proportion of X chromosome markers in the cell line carrying 3X chromosomes that have copy number values below the same threshold. The curve is generated by calculating these values at many different thresholds.

Figure 5. ROC analysis for 2-marker smoothing

The same ROC analysis as described in Figure 2 is performed here using the same set of markers (~40K) as well as using a larger number of markers (~48K). The ROC analysis was also performed using 2-marker smoothing. In this case the smoothing was done for two random markers. If we assume that the performance of individual markers is not correlated with their position (i.e. markers close together are likely to have similar performance) then this should be an accurate reflection of the resultant performance with adjacent markers smoothing. We note that at the lower false positive rate for the 2-markers smoothed data, the curve is not smooth given low statistics.

Figure 6. Amplification in MCF7

A) The X axis shows the markers in a genomic order, with each chromosome uniquely colored. The Y chromosome depicts the measured copy number for each marker in $\log_2$ (the log scale is used given the high dynamic range). The arrow depicts the position of the locus that was also analyzed by real time PCR. B) Focused view around the amplification site that was checked with real time PCR. As can be seen there are several sites of amplifications of different levels. The black bar identifies the region for which average copy number was calculated.

Figure 7. Estimation of copy number of the spikes

The X-axis shows the expected copy number (in $\log_2$) for the individual spiked in PCR fragments, and the Y-axis shows the observed copy number for the same spiked in fragments. The linear fit (r^2=0.82) is only for spikes with expected copy number <64 (2^6) because of the clear saturation above that point.

Figure 8. Allele Copy Number distributions and reproducibility

Copy number measurements for tumor sample 47 (fairly normal genome content) is shown in panel A with genotypes AA colored red, AB colored blue, and BB colored green. Allele copy number measurements for tumor 45 (replicate 1) are plotted in panels B and C. In panel B, genotypes derived from replicate 1 are colored AA red, AB blue, and BB green. Panel C is the genotypes from replicate 2 in the same color scale.

Figure 9. Visualization of individual copy number measurements without matched normal samples

Panels A-C show copy number measurements for tumor 48 in genome order from chromosome 1 on the left to chromosome 22 and X on the right. Data are segregated by higher and lower copy number estimates and by homozygosity or heterozygosity. Blue and orange data points are the higher allele copy measurement while green and red data points are the lower copy number measurements. Blue and red data points are homozygous alleles while orange and green are heterozygous alleles. Panel A shows the entire genome. Panel B shows chromosome 1 through the first 100Mbp of chromosome 5. Panel C shows chromosome 1 and the first 50Mbp of chromosome 2. Panel C shows key features of ACN data. An amplification is seen near position 5e7. An extra copy of

1q is seen between ~1.5e8 and 2.5e8. A deletion of 1 copy is seen on the p arm of chromosome 2 between ~2.5e8 and 3e8 (observed in panel B as a complete loss of one copy of chromosome 2). Panel D shows a small section of chromosome 5 from tumor 44. One chromosome is at copy number 0.5 across this region, which indicates a loss of that chromosome. The black arrow shows a region at total copy number 2, which likely includes reduplication of the lost chromosome in the tumor. The red arrow shows a region where both alleles are at copy number 0.5 suggesting a complete deletion. The green arrow shows copy number 1 for the yellow alleles

Tables

Table 1

Sensitivity at 50% Specificity

	1 Marker	2 Markers
40K (75% of data)	1.7E-03	4.0E-05
48K (90% of data)	2.7E-03	7.1E-05

Table 2

Expected vs. Measured Copy Number

Expected copy number	Measured copy number	Relative standard deviation
1 (9)	1.055	0.14
2 (15)	1.997	0.12
3 (2)	3.104	0.11
4 (2)	3.981	0.10
5 (2)	4.956	0.10

Table 3

Allele Copy Number in Spiked Samples

Copy_A	Copy_B
199.2	1.3
184.8	1.1
169.4	1.5
141.8	1.3
139.7	0.8
105.4	1.0
84.6	1.0
80.2	0.9
73.8	0.8

73.0	0.9
70.6	1.1
64.5	1.0
60.8	1.0
59.8	1.3
57.4	0.8
54.6	1.1
52.8	1.0
43.3	0.8
39.2	1.1
38.8	0.9
33.4	0.8
27.5	0.8
25.7	0.9
18.7	1.2
14.3	0.7
12.9	0.9
11.3	1.0

Table 4

Genotyping disagreements between replicated samples

Sample	Discordant Calls	Total Calls	Discordant Rate
44	643	50236	1.3%
45	419	50260	0.8%
46	271	50250	0.5%
47	258	50244	0.5%
48	393	50242	0.8%

Additional Files

Supplementary file 1. XLS. Allelic Copy Number of breast tumors. Replicated CAN data for five breast cancers. Data are filtered to use only high quality MIPs (90% or greater call rate, less than 12% copy number variation).

Figure 1

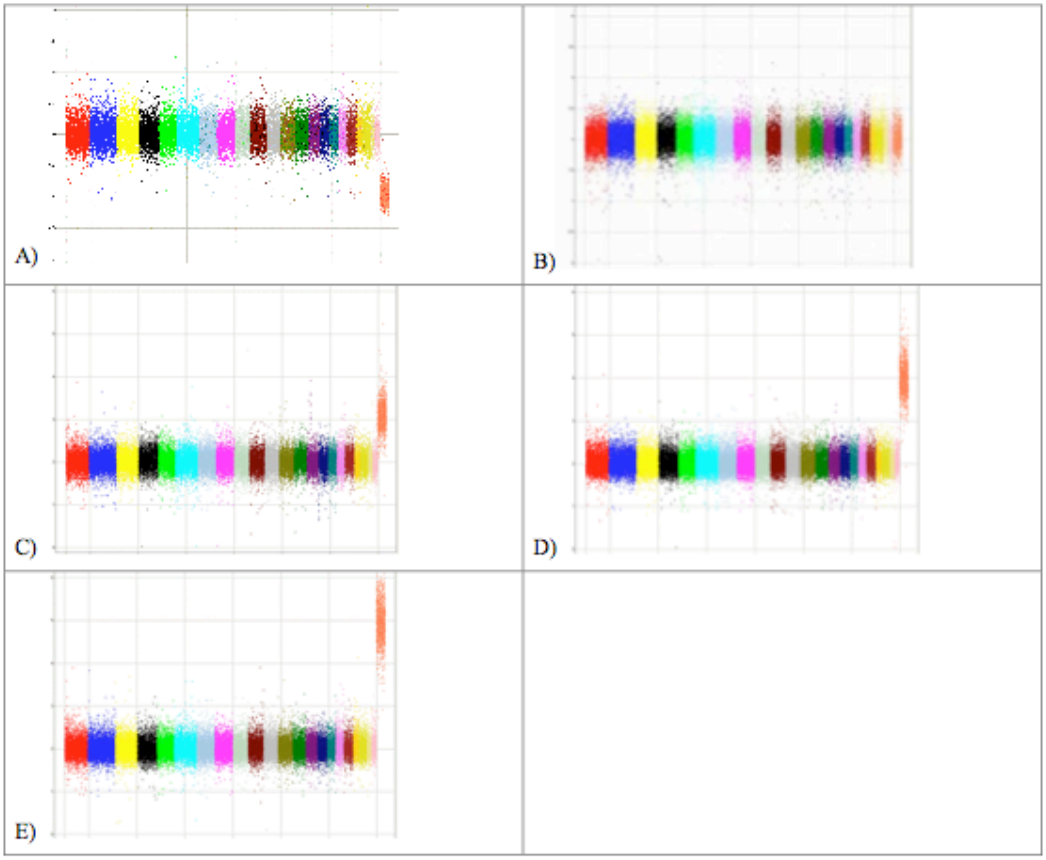

Figure 2

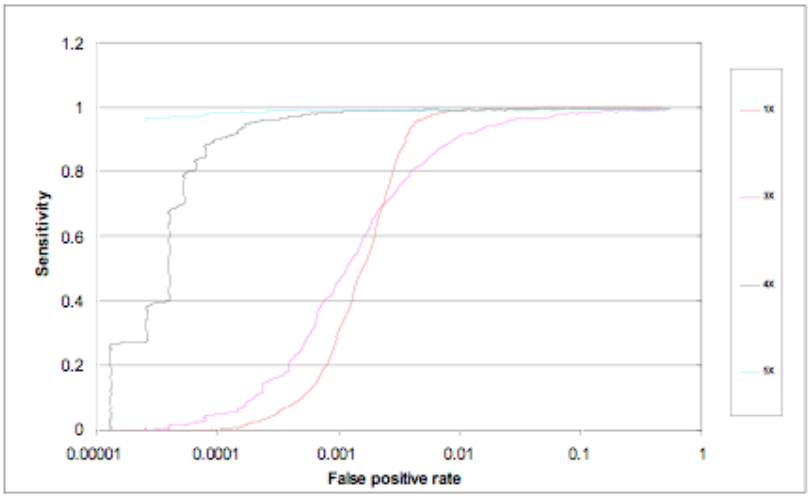
1.2
1
0.8
0.6
0.4
0.2
0
Sensitivity
0.00001
0.0001
0.001
0.01
0.1
1
False positive rate

Figure 3

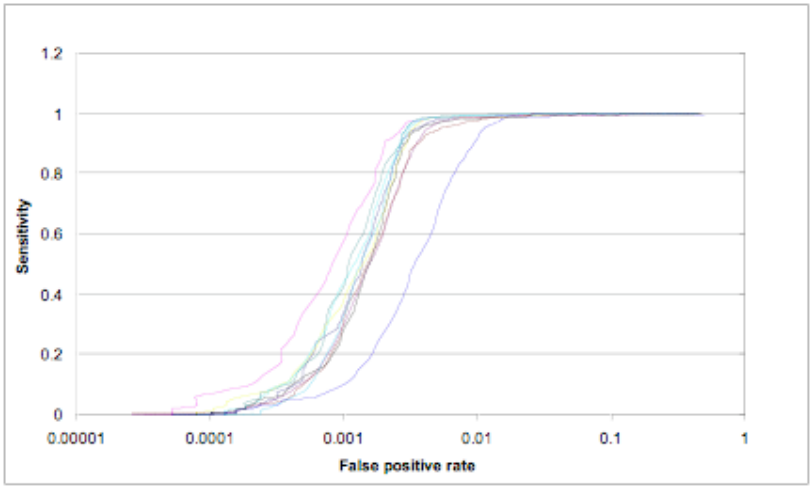
1.2
1
0.8
0.6
0.4
0.2
0
Sensitivity
0.00001
0.0001
0.001
0.01
0.1
1
False positive rate

Figure 4

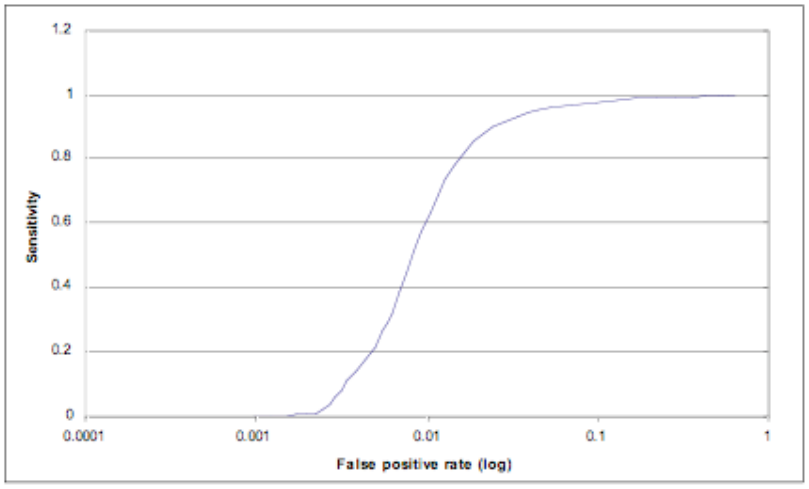
1.2
1
0.8
0.6
0.4
0.2
0
Sensitivity
0.0001
0.001
0.01
0.1
1
False positive rate (log)

Figure 5

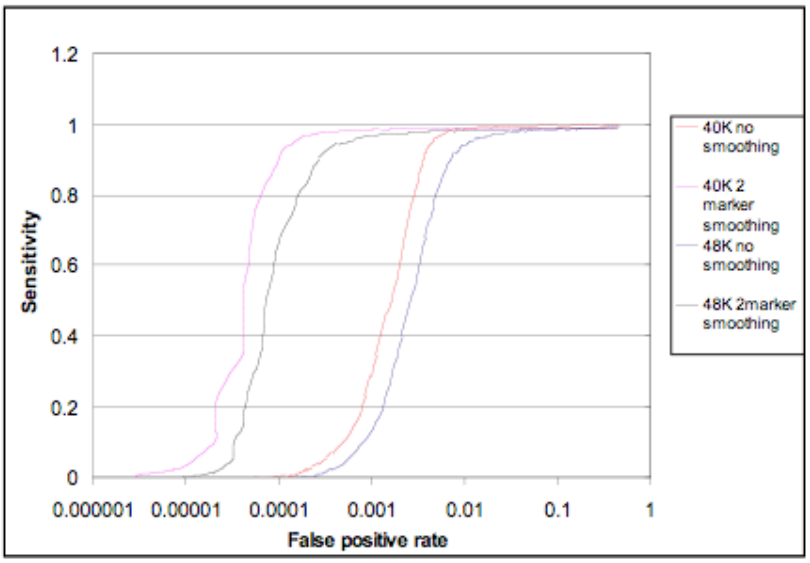
1.2
1
0.8
0.6
0.4
0.2
0
Sensitivity
0.000001
0.00001
0.0001
0.001
0.01
0.1
1
False positive rate
40K no smoothing
40K 2 marker smoothing
48K no smoothing
48K 2marker smoothing

Figure 6

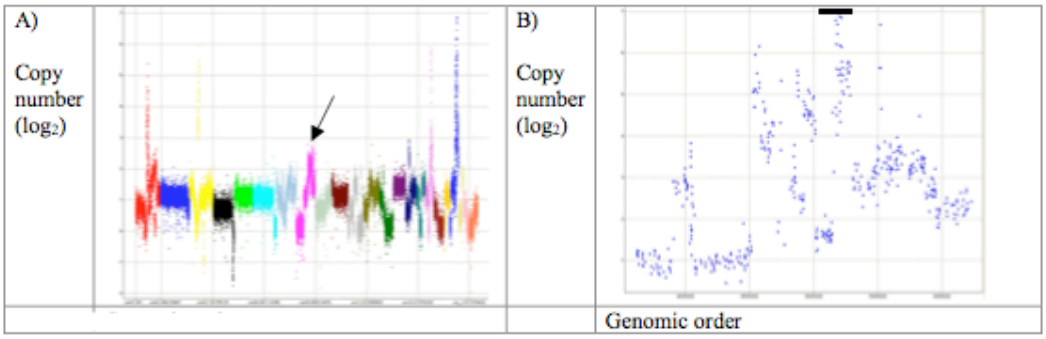

Figure 7

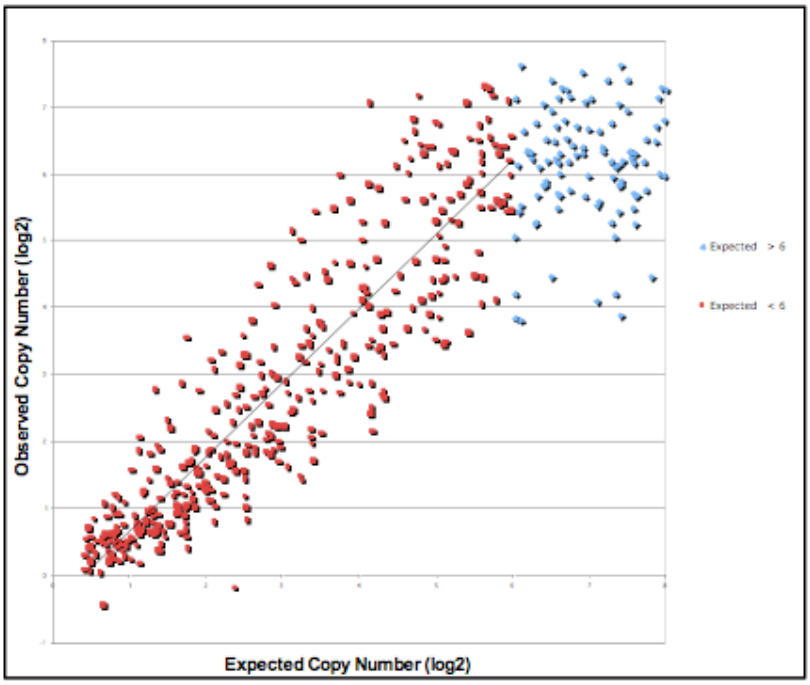
Observed Copy Number (log2)
Expected Copy Number (log2)
Expected > 6
Expected < 6

Figure 8

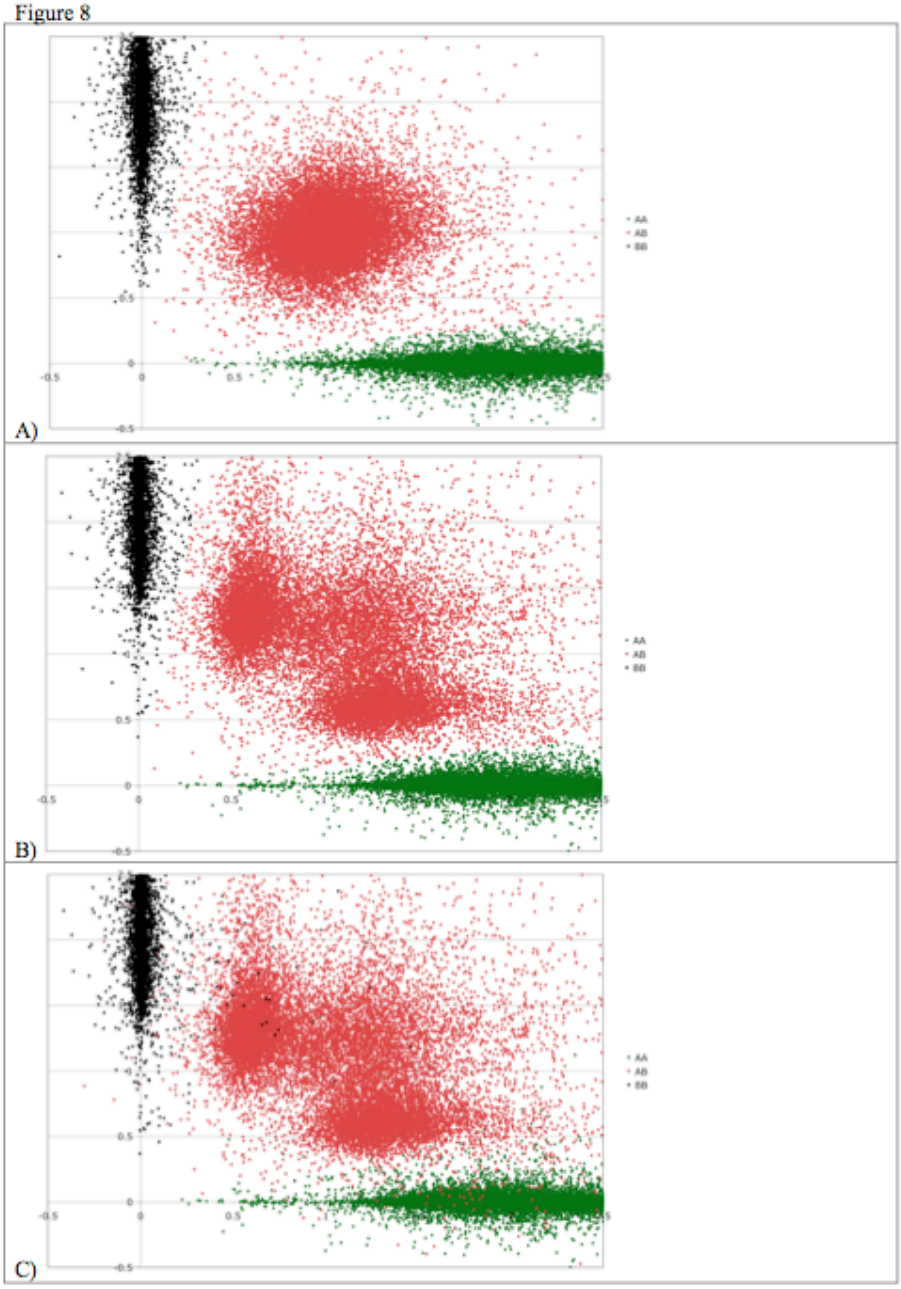

Figure 9